Evaluation of an Automatic F-Structure Annotation Algorithm against the PARC 700 Dependency Bank

Michael Burke, Aoife Cahill, Ruth O'Donovan, Josef van Genabith, Andy Way
National Centre for Language Technology and School of Computing,
Dublin City University, Dublin, Ireland
{mburke,acahill,rodonovan,josef,away}@computing.dcu.ie

— *Abstract* —

An automatic method for annotating the Penn-II Treebank with high-level Lexical Functional Grammar f-structure representations is described in [1, 2, 3, 8]. Annotation coverage is near complete with 99.83% of the 48K Penn-II sentences receiving a single, connected f-structure. The annotation algorithm and the automatically produced f-structures are the basis for the automatic acquisition of wide-coverage and robust PCFG-based approximations of LFG grammars ([2, 3]) and for the induction of LFG semantic forms ([8]). Therefore the quality of the annotation algorithm and the f-structures it generates is extremely important. To date, annotation quality has been measured in terms of precision and recall against a set of manually constructed, gold-standard f-structures for 105 randomly selected sentences from Section 23 of the WSJ part of Penn-II. The algorithm currently achieves an f-score of 96.3% for complete f-structures and 93.6% for preds-only f-structures using the evaluation methodology and software presented in [5] and [9].

There are a number of problems with evaluating against a gold standard of this size, most notably that of overfitting. There is a risk of assuming that the gold standard is a complete and balanced representation of the linguistic phenomena in a language and basing design decisions on this. It is, therefore, preferable to evaluate against an independently constructed, more extensive, external standard. Although the 105 gold standard f-structures are publicly available[1], a larger well-established external standard can provide a more widely-recognised benchmark against which the quality of the algorithm can be evaluated. For these reasons, we present an evaluation of the annotation algorithm of [1, 2, 3, 8] against the PARC 700 Dependency Bank [7]. The PARC 700 comprises 700 randomly selected sentences from Section 23 of the WSJ section of Penn-II which were parsed by a hand-coded, deep LFG, converted to dependency format (triples) and manually corrected and extended. We use the automatic annotation algorithm of [1, 2, 3, 8] to generate f-structures for those 700 Penn-II trees and also a subset of 560 as outlined in [6].

Evaluation against an external standard is a non-trivial and time-consuming task, in this case due primarily to systematic differences in linguistic analysis, feature geometry and nomenclature. In order to carry out the evaluation we developed conversion software to automatically deal with systematic differences (Figure 1).

[1] Available on http://www.computing.dcu.ie/research/nclt/gold105.txt.

Before annotating Penn-II trees we deal with named entity recognition. The PARC 700 analyses certain names (e.g. 'Merrill Lynch') as complex predicates while the annotation algorithm analyses the same string fully parsed as a head ('Lynch') modified by an adjunct ('Merrill'). Our pre-processing named entity recognition component identifies and tags named entities in the Penn-II trees. The trees are then annotated by the f-structure annotation algorithm and passed through three post-processing components.

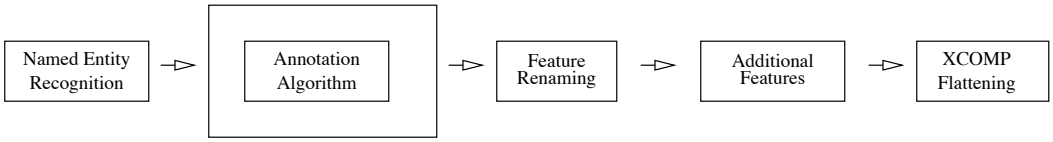

Figure 1: Conversion Software

A significant number of feature names differ between the PARC 700 gold standard and the automatically generated f-structures. The first post-processing component implements a mapping to establish common or merged feature names. A number of features in the PARC 700 are not computed within the annotation algorithm, e.g. `stmt_type`. Some common features have differing value ranges, e.g. the `adegree` value `positive` occurs frequently in the PARC 700, while our f-structures only ever have `comparative` and `superlative` as values. The second post-processing component systematically annotates the trees with the missing features and values.

One key difference in the f-structures produced by the annotation algorithm and those in the PARC 700 is the representation of tense and aspect information. While the annotation algorithm in [1, 2, 3, 8] uses a system of cascading XCOMPs to encode this information at f-structure level, the same details are represented in the PARC 700 using flat f-structures and tense and aspect features. To cope with this, we automatically flatten the f-structures produced by the annotation algorithm, essentially stripping the outer XCOMPs while carrying over remaining f-structure information from each level of embedding.

	Entire PARC 700	**560 sentence subset**
Precision	88.1	88.06
Recall	85.3	85.27
F-Score	86.68	86.64

Table 1: Results of F-Structure Evaluation against the PARC 700 using the feature set from [6]

Our work on the conversion process is ongoing. The automatic f-structure annotation algorithm in [1, 2, 3, 8] currently achieves an f-score of 86.64% against the PARC 700 (Table 1). We have used the same evaluation procedure to evaluate the output generated by our automatically induced PCFG-based LFG approximations against the PARC 700. Currently, our best grammar performs at 80.9% precision and 76.74% recall with an f-score of 78.77%.

References

[1] Cahill, A., M. McCarthy, M. Burke, R. O'Donovan, J. van Genabith and A. Way. 2004. Evaluating Automatic F-Structure Annotation for the Penn-II Treebank. In *Journal of Research on Language and Computation*, Kluwer Academic Publishers (in print).

[2] Cahill, A., M. McCarthy, J. van Genabith and A. Way (2002b): 'Parsing Text with a PCFG derived from Penn-II with an Automatic F-Structure Annotation Procedure', in M. Butt and T. Holloway-King (eds.) *Proceedings of the Seventh International Conference on Lexical-Functional Grammar*, CSLI Publications, Stanford, CA., pp.76–95.

[3] Cahill, A., M. Burke, R. O'Donovan, J. van Genabith and A. Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. To appear in *Proceedings of 42nd Conference of the Association for Computational Linguistics*, Barcelona.

[4] Carroll, J., G. Minnon, and T. Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC)*, Bergen, Norway.

[5] Crouch, R., R. Kaplan, T. King and S. Riezler. 2002. A comparison of evaluation metrics for a broad coverage parser. In *Beyond PARSEVAL workshop* at 3rd Int. Conference on Language Resources and Evaluation (LREC'02), Las Palmas.

[6] Kaplan, R., S. Riezler, T. King, J. Maxwell, A. Vasserman and R. Crouch. 2004. In *Proceedings of the Human Language Technology Conference and the 4th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*, Boston, MA.

[7] King, T. H., R. Crouch, S. Riezler, M. Dalrymple, and R. M. Kaplan (2003). The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest.

[8] O'Donovan, R., M. Burke, A. Cahill, J. van Genabith and A. Way. 2004. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II Treebank. To appear in *Proceedings of 42nd Conference of the Association for Computational Linguistics*, Barcelona.

[9] Riezler, S., R. Kaplan, T. King, M. Johnson, R. Crouch and J. Maxwell III. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of 40th Conference of the Association for Computational Linguistics*, Philadelphia, PA.